

W1200

DISK ARRAY CONTROLLER

Patent number: JP10333836
Publication date: 1998-12-18
Inventor: FUJIMOTO KAZUHISA; TANAKA ATSUSHI; ODAWARA HIROAKI
Applicant: HITACHI LTD
Classification:
- international: G06F3/06; G06F3/06; G11B20/10; G11B20/18; G11B20/18
- european:
Application number: JP19970139656 19970529
Priority number(s):

[View INPADOC patent family](#)

Abstract of JP10333836

PROBLEM TO BE SOLVED: To reduce the communication load on a shared memory part, and to improve the total throughput of a disk array controller, by collecting plural transmission data having the same sending destination in a single composite packet.

SOLUTION: The control information sent from a microprocessor 100 or the data sent from a data transmission/reception control part 110, which performs the transmission/reception of data to a host computer are sent to a communication controller 140 and then stored in the buffers 130 and 131 prepared for every sending destination. When plural transmission data having the same sending destination are stored, a packet generation part 120 collects these data in a composite packet and sends it to a desired shared memory part. Then, the composite packet sent from the shared memory part in a composite packet are decomposed into plural transmission data at a packet decomposition part 125 and stored in a buffer 135. These decomposed data are sent to a microprocessor 106 or the control part 110 via the controller 140.

(19)日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11)特許出願公開番号

特開平10-333836

(43)公開日 平成10年(1998)12月18日

(51)Int.Cl. ⁸	識別記号	F I		
G 0 6 F 3/06	3 0 1	G 0 6 F 3/06	3 0 1 R	
	5 4 0		5 4 0	
G 1 1 B 20/10		G 1 1 B 20/10	D	
20/18	5 7 0	20/18	5 7 0 Z	
	5 7 2		5 7 2 F	
審査請求 未請求 請求項の数 5 O L (全 7 頁)				

(21)出願番号 特願平9-139656

(22)出願日 平成9年(1997)5月29日

(71)出願人 000005108

株式会社日立製作所

東京都千代田区神田駿河台四丁目6番地

(72)発明者 藤本 和久

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 田中 淳

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

(72)発明者 小田原 宏明

東京都国分寺市東恋ヶ窪一丁目280番地

株式会社日立製作所中央研究所内

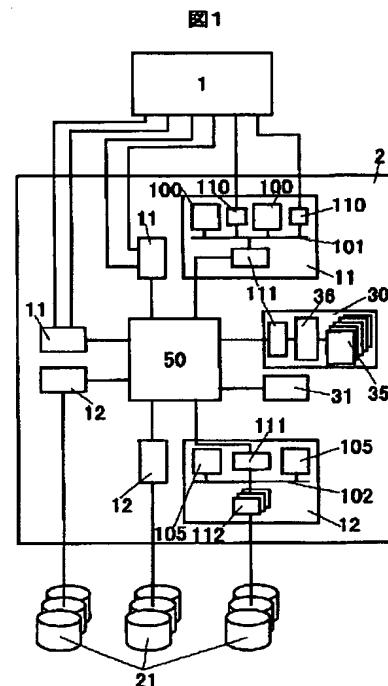
(74)代理人 弁理士 小川 勝男

(54)【発明の名称】 ディスクアレイ制御装置

(57)【要約】

【課題】ディスクアレイ装置において、高いI/Oスループット性能を実現する。

【解決手段】ホストコンピュータ1とのインターフェース部11と、複数の磁気ディスク装置21とのインターフェース部12と、共有メモリ部30、31内の、スイッチとの通信制御部111において、バッファ内に格納された送出先が同一の複数の送信データがあるときは、それらを1つの複合パケットにまとめて送出する。



【特許請求の範囲】

【請求項1】少なくとも、ホストコンピュータとの1つのインターフェース部と、複数の磁気ディスク装置との1つのインターフェース部と、データ及び制御情報を格納する物理的に独立した複数の共有メモリ部から成り、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記複数の共有メモリ部が、スイッチを用いた相互結合網によって結合されたディスクアレイ制御装置であって、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部において、バッファ内に格納された送出先が同一の複数の送信データを1つの複合パケットにまとめて送出することを特徴とするディスクアレイ制御装置。

【請求項2】請求項1記載のディスクアレイ制御装置であって、前記共有メモリ部内のメモリ制御部において、該共有メモリ部に送出された前記複数の送信データから成る複合パケット内の複数の送信データを並列に処理することを特徴とするディスクアレイ制御装置。

【請求項3】請求項1または2において、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部内のバッファが、それぞれの送信データの送出先ごとに、物理的あるいは論理的に分割されていることを特徴とするディスクアレイ制御装置。

【請求項4】請求項1から3において、前記ホストコンピュータとのインターフェース部、及び前記複数の磁気ディスク装置とのインターフェース部が、該インターフェース部内の処理を分散して行う複数のマイクロプロセッサから成ることを特徴とするディスクアレイ制御装置。

【請求項5】請求項1から4において、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部内のバッファからの送信データ送出処理の際、該バッファ内に、送出先が同一の送信データが少なくとも2つ以上格納されている場合にのみ、該送信データ群を1つの複合パケットにまとめて送出することを特徴とするディスクアレイ制御装置。

【発明の詳細な説明】**【0001】**

【発明の属する技術分野】本発明は、データを分割して複数の磁気ディスク装置に格納するディスクアレイ装置の制御装置に関する。

【0002】

【従来の技術】コンピュータの主記憶のI/O性能に比べて、2次記憶装置として用いられる磁気ディスク装置を使ったサブシステムのI/O性能は3～4桁程度小さく、

従来からこの差を縮めること、すなわちサブシステムのI/O性能を向上する努力が各所でなされている。サブシステムのI/O性能を向上させるための1つの方法として、複数の磁気ディスク装置でサブシステムを構成し、データを分割して複数の磁気ディスク装置に格納する手段、いわゆるディスクアレイと呼ばれるシステムが知られている。

【0003】このディスクアレイでは、複数の磁気ディスク装置を並列に動作させてデータの入出力を行うため、I/O性能が向上する。しかし、複数の磁気ディスク装置への書き込み及び読み出し動作を並列に処理するため、制御が複雑で、それに要するオーバーヘッドが大きいという問題がある。

【0004】図2に主にメインフレーム向けの大型ディスクアレイ装置の制御装置3を示す。この制御装置3は一般的に、ホストコンピュータ1とのインターフェース部13、複数の磁気ディスク装置21とのインターフェース部14、データ及び制御情報を一時格納する共有メモリ部32が、共有バス60を介して繋がる構成をとっている。

【0005】ディスクアレイ装置のI/Oスループット性能の伸びは大きく、それに対応するため、上記インターフェース部等の処理性能を向上させる必要がある。これら処理性能の向上に伴って、共有バス60の利用率が飽和状態となり、それが原因でスループット性能が制限されている。そこで、共有バス60のスループットを上げるための努力がなされているが、装置の構成上、バス幅、駆動周波数等を改善することは難しく、スループットの向上にも限界がある。

【0006】そのため、共有バス60に代わって、スイッチを用いた相互結合網を介して上記インターフェース部等を繋ぐことが考えられている。この方法では、相互接続された個々のバスのスループットは、共有バスの数分の1であるが、相互接続された2点間には複数のバスが存在するため、負荷が分散され、スループットの向上が可能となる。

【0007】

【発明が解決しようとする課題】スイッチを用いた相互結合網では、多対多の通信が一般的であるが、ディスクアレイの制御装置では、図2に示す複数のインターフェース部13、14と共有メモリ部32間の通信が大部分を占める。したがって、負荷分散のため、共有メモリ部を物理的に複数の分割することが考えられている。しかしながら、共有メモリ部の分割数にも限界があり、均等な負荷分散を行うことは難しい。したがって、ディスクアレイ装置においては、スイッチを用いた相互結合網が持つ本来のスループット性能を引き出すことは難しく、スループット向上にも限界がある。

【0008】本発明の目的は上述の課題を解消し、I/Oスループット性能の高いディスクアレイ装置を提供す

ることにある。

【0009】

【課題を解決するための手段】上記目的は、少なくとも、ホストコンピュータとの1つのインターフェース部と、複数の磁気ディスク装置との1つのインターフェース部と、データ及び制御情報を格納する物理的に独立した複数の共有メモリ部から成り、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記複数の共有メモリ部が、スイッチを用いた相互結合網によって結合されたディスクアレイ制御装置であって、前記ホストコンピュータとのインターフェース部と、前記複数の磁気ディスク装置とのインターフェース部と、前記共有メモリ部内の、スイッチとのインターフェース部において、バッファ内に格納された送出先が同一の複数の送信データを1つの複合パケットにまとめて送出するディスクアレイ制御装置によって達成される。

【0010】すなわち、送出先が同一の複数の送信データを1つの複合パケットにまとめることにより、複数回の通信オーバーヘッドが1回に減るため、1つのパケット長は長くなるが全体の通信量は減る。したがって、共有メモリ部への通信負荷が減り、制御装置全体のスループットの向上が可能となる。

【0011】また、共有メモリ部内のメモリ制御部で、その共有メモリ部に送出された複数の送信データから成る複合パケット内の複数の送信データを並列に処理することによって、共有メモリ部のスループットの向上が可能となる。これにより、制御装置全体のスループットがさらに向上する。

【0012】また、ホストコンピュータとのインターフェース部と、複数の磁気ディスク装置とのインターフェース部と、共有メモリ部内の、スイッチとのインターフェース部内のバッファを、送信データそれぞれの送出先ごとに物理的あるいは論理的に分割することにより、送出先が同一の送信データを1つの複合パケットにまとめる処理が簡単になる。これにより、スイッチとのインターフェース部での処理オーバーヘッドの削減が可能となる。

【0013】ホストコンピュータとのインターフェース部、及び複数の磁気ディスク装置とのインターフェース部が、そのインターフェース部内の処理を分散して行う複数のマイクロプロセッサから成っている場合、複数のマイクロプロセッサが同一の共有メモリ部にほぼ同時にアクセスすることが発生するため、スイッチとのインターフェース部のバッファ内に、送出先が同一の送信データが複数存在する頻度が高くなる。したがって、この場合、上記手段は特に有効となる。

【0014】スイッチとのインターフェース部内のバッファからの、送信データ送出処理の際、該バッファ内に送出先が同一の送信データがない場合に、送出先が同一

の送信データ発生を待つと、待ち時間が長くなった場合に、制御装置全体の応答性能が悪くなる可能性がある。そのため、送出先が同一の送信データが少なくとも2つ以上格納されている場合にのみ、該送信データ群を1つの複合パケットにまとめて送出することが望ましい。

【0015】

【発明の実施の形態】

（実施例1）図1に、本発明の制御装置の一実施例を示す。ディスクアレイ制御装置2内の、ホストコンピュータ1とのインターフェース部11と、複数の磁気ディスク装置21とのインターフェース部12と、2つの共有メモリ部30、31は、スイッチを用いた相互結合網50を介して結合されている。

【0016】スイッチを用いた相互結合網50は、複数のポートを有する少なくとも1つのスイッチから成る。各スイッチのポートには、他のスイッチ、あるいはディスクアレイ制御装置内の各インターフェース部及び共有メモリ部が繋がる。本実施例では、図7に示すように入出力にそれぞれ5つのポートを有するスイッチ40を4つ、相互に結合した（完全結合網と呼ぶトポロジー）相互結合網50を用いる。図では、入出力ポートをまとめて1つの線として示している。他のトポロジーの相互結合網を用いても本実施例を実施する上で問題はない。

【0017】ホストコンピュータ1とのインターフェース部11内では、2つの制御用マイクロプロセッサ100、ホストコンピュータとの2つのデータ送受信制御部110、及びインターフェース部-共有メモリ部間の通信制御部111が共有バス101を介して接続されている。

【0018】複数の磁気ディスク装置21とのインターフェース部12内では、2つの制御用マイクロプロセッサ105、複数の磁気ディスク装置21とのデータ送受信制御部112、及びインターフェース部-共有メモリ部間の通信制御部111が共有バス102を介して接続されている。

【0019】共有メモリ部30、31内では、インターフェース部-共有メモリ部間の通信制御部111を介して、メモリ制御部36、メモリモジュール35が接続されている。

【0020】ホストコンピュータ1からディスクアレイ制御装置2へのデータの読み出し要求は、ホストコンピュータとのデータの送受信制御部110を通して、マイクロプロセッサ100に伝える。マイクロプロセッサ100は、共有メモリ30、31内に要求データが有るかどうかを確認するため、インターフェース部-共有メモリ部間の通信制御部111を通して通信を行う。要求データが存在する場合、マイクロプロセッサ100は、共有メモリからの読み出し処理を実行する。要求データが存在しない場合、マイクロプロセッサ100は、磁気ディスク装置21から共有メモリ部30あるいは31への要求データの転送命令を、共有メモリ部30あるいは31

1, 目的ポートNo. (送出先のID番号) 262, ルーティング情報 (途中経由するスイッチの番号) 263 から成る。これらの内容を經由スイッチ内で参照し、パケットを目的のポートまで届ける。この方法以外に、ルーティング情報263無しに、スイッチ内で次に経由するスイッチを動的に割り当てることによって、パケットを目的ポートへ届ける方法もある。この方法によっても、本発明を実施する上で問題はない。

【0031】図5は、複数の磁気ディスク装置21とのインターフェース部12内の、スイッチとの通信制御部111の構成を示している。インターフェース部12内では、2つの制御用マイクロプロセッサ105, 複数の磁気ディスク装置21とのデータ送受信制御部112、及びインターフェース部-共有メモリ部間の通信制御部111が共有バス102を介して接続されている。このインターフェース部12においても、図3, 図4において説明した方法と同様の制御を、スイッチとの通信制御部111において行う。

【0032】図6は、共有メモリ部30, 31内の、スイッチとの通信制御部111の構成を示している。共有メモリ部30, 31内では、インターフェース部-共有メモリ部間の通信制御部111を介して、メモリ制御部36, メモリモジュール35が接続されている。この共有メモリ部30, 31においても、図3, 図4において説明した方法と同様の制御を、スイッチとの通信制御部111において行う。

【0033】以上述べたように、送出先が同一の複数の送信データを1つの複合パケットにまとめることにより、複数回の通信オーバーヘッドが1回に減るため、1つのパケット長は長くなるが全体の通信量は減る。したがって、共有メモリ部への通信負荷の削減が可能となる。

【0034】ここでは、ホストコンピュータ1とのインターフェース部11、及び複数の磁気ディスク装置21とのインターフェース部12が、そのインターフェース部内の処理を分散して行う2つのマイクロプロセッサ100あるいは105から成っている場合について説明したが、インターフェース部11あるいは12が、1つのマイクロプロセッサから成っている場合も、本発明を実施する上で問題はない。しかしながら、インターフェース部11あるいは12が複数のマイクロプロセッサから成っている場合、それらが同一の共有メモリ部にほぼ同時にアクセスすることが発生するため、スイッチとのインターフェース部のバッファ内に、送出先が同一の送信データが複数存在する頻度が高くなる。したがって、この場合、本発明は特に有効である。

【0035】図8は、スイッチによる相互結合網を用いた従来のディスクアレイ制御装置と本発明のディスクアレイ装置のスループット性能を計算によって求め、比較した結果を示している。ホストからの負荷の条件は、ベ

ンチマークプログラムPAI-I/O-Driverのzero-Localityでの負荷条件(Read:Write比が1:1で、キャッシュヒット率が0%)とした。ディスクアレイ制御装置は、ホストコンピュータとのインターフェース部11のパッケージ8枚と複数の磁気ディスク21とのインターフェース部12のパッケージが16枚、共有メモリ部のパッケージが8枚から構成されているとした。

【0036】図では、縦軸に共有メモリへのアクセス時間を、限界のアクセス時間を1とした相対値で示している。また、横軸にスループットを、従来の限界値を1とした相対値で示している。図からわかるように、従来に比べて本発明のディスクアレイ制御装置では、スループットが約26%向上する。

【0037】ディスクアレイ制御装置2では、各インターフェース部11, 12と共有メモリ部30, 31との間の通信が中心となってホストコンピュータ1からのI/O要求の処理が行われる。したがって、ホストコンピュータからのI/O負荷が増加するにつれて、上記の通信量が増大する。その結果、スイッチを用いた相互結合網の利用率が飽和状態となり、スループット性能が制限される。しかしながら、本実施例によれば、共有メモリ部への通信負荷が削減でき、スイッチを用いた相互結合網全体のスループットが上がるため、ディスクアレイ制御装置2全体のスループット向上が可能となる。

【0038】(実施例2) 本発明の他の一実施例を示す。実施例1で述べたディスクアレイ制御装置2において、共有メモリ部30, 31内のメモリ制御部36で、その共有メモリ部に送出された複数の送信データから成る複合パケット内の複数の送信データを並列に処理する機能を持たせる。すなわち、図9に示すように、複数のメモリモジュール35へのバスを複数設け、それぞれのバス毎に制御コントローラ37を設けることによって、複数の送信データを並列に処理する。

【0039】本実施例によれば、共有メモリ部のスループットが上がるため、ディスクアレイ制御装置2全体のスループットをさらに向上することが可能となる。

【0040】(実施例3) 本発明の他の一実施例を示す。実施例1または2で述べたディスクアレイ制御装置2において、スイッチとの通信制御部111のバッファ内からの送信データ送出処理の際、そのバッファ内に送出先が同一の送信データがない場合に送出先が同一の送信データ発生を待つと、待ち時間が長くなった場合に制御装置全体の応答時間性能が悪くなる場合があると考えられる。そこで、スイッチとの通信制御部111で、送出先が同一の送信データが少なくとも2つ以上格納されている場合にのみ、該送信データ群を1つの複合パケットにまとめて送出する。送信データが1つしかない場合には、そのみをパケットとして送出するという制御を行う。

【0041】本実施例によれば、複合パケットによる共

有メモリ部との通信を行った場合のディスクアレイ制御装置全体の応答時間性能の低下を防ぐことが可能となる。

【0042】

【発明の効果】本発明によれば、スイッチを用いた相互結合網内において共有メモリ部への負荷が減るため、そのスループットが向上する。それによって、ディスクアレイ制御装置の各インターフェース部と共有メモリ部間の通信の飽和が解消され、装置全体のスループット性能が向上する。

【図面の簡単な説明】

【図1】本発明の一実施例のディスクアレイ制御装置の構成を示すブロック図。

【図2】従来のディスクアレイ制御装置の構成を示すブロック図。

【図3】図1のホストコンピュータとスイッチとのインターフェース部の構成を示すブロック図。

【図4】本発明の一実施例のデータのフォーマットを示す説明図。

【図5】図1の磁気ディスク装置とスイッチとのインターフェース部の構成を示すブロック図。

【図6】図1の共有メモリとスイッチとのインターフェ

ース部の構成を示すブロック図。

【図7】図1のスイッチを用いた相互結合網の構成を示すブロック図。

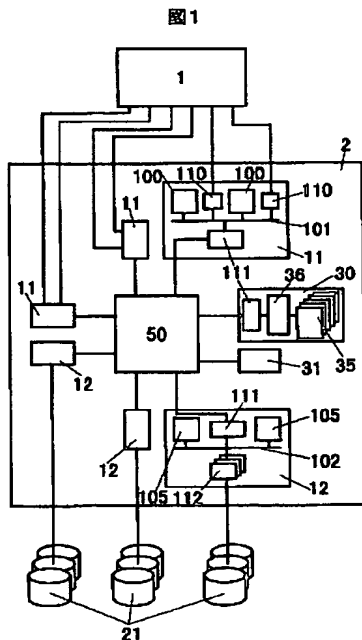
【図8】従来と本発明によるディスクアレイ制御装置のスループット性能の予測結果を比較した図。

【図9】本発明の他の実施例における共有メモリ部の構成を示すブロック図。

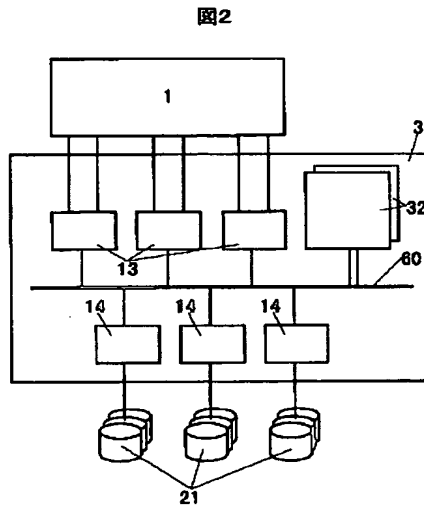
【符号の説明】

1…ホストコンピュータ、2…ディスクアレイ制御装置、11…ホストコンピュータとのインターフェース部、12…磁気ディスク装置とのインターフェース部、21…磁気ディスク装置、30…共有メモリ部、35…メモリモジュール、36…メモリ制御部、50…スイッチを用いた相互結合網、100…マイクロプロセッサ、101…共有バス、102…共有バス、105…マイクロプロセッサ、110…ホストコンピュータとのデータの送受信制御部、111…通信制御部、112…磁気ディスク装置とのデータ送受信制御部、120…パケット生成部、125…パケット分解部、130、131…送信先別バッファ、135…バッファ、140…通信制御コントローラ。

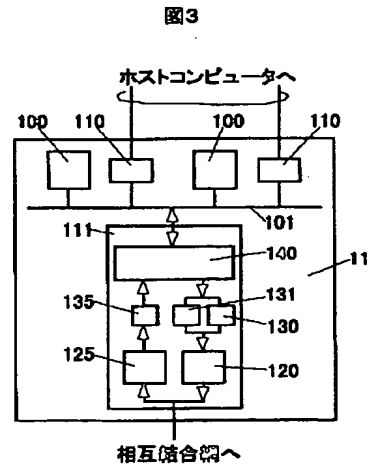
【図1】



【図2】

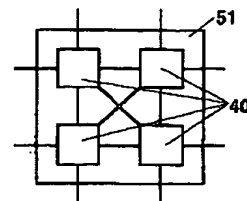


【図3】

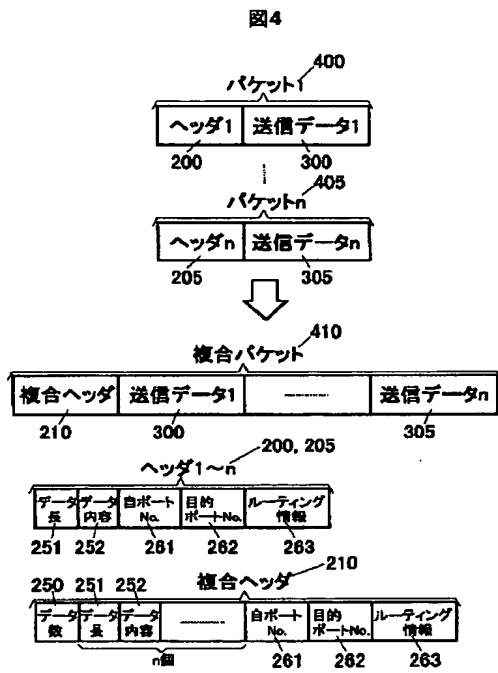


【図7】

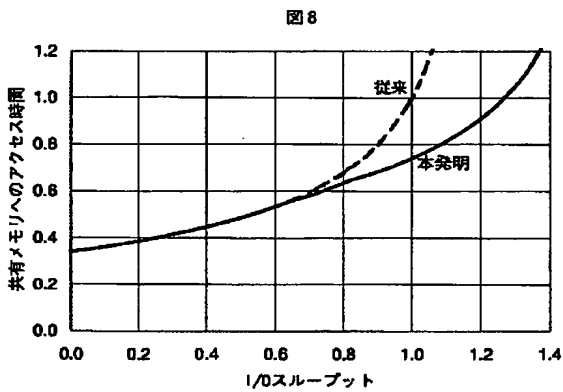
図7



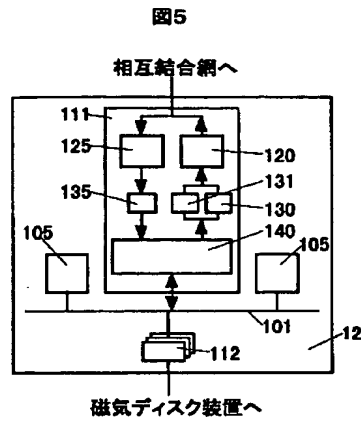
【図4】



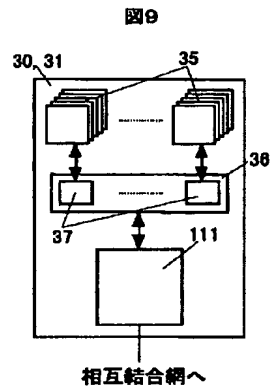
【図8】



【図5】



【図9】





Creation date: 02-11-2004
Indexing Officer: KSAM - KONA SAM
Team: OIPEScanning
Dossier: 10700485 ✓

Legal Date: 01-09-2004 ✓

No.	Doccode	Number of pages
1	IDS	9
2	NPL	4
3	NPL	4
4	NPL	7
5	NPL	7

Total number of pages: 31

Remarks:

Order of re-scan issued on